Ontology Population via NLP techniques in Risk Management

Jawad Makki, Anne-Marie Alquier, and Violaine Prince

Abstract— One of the challenging tasks in the context of Ontological Engineering is to automatically or semi-automatically support the process of Ontology Learning and Ontology Population from semi-structured documents (texts). In this paper we describe a Semi-Automatic Ontology Instantiation method from natural language text, in the domain of Risk Management. This method is composed from three steps 1) Annotation with part-of-speech tags, 2) Semantic Relation Instances Extraction, 3) Ontology instantiation process. It's based on combined NLP techniques using human intervention between steps 2 and 3 for control and validation. Since it heavily relies on linguistic knowledge it is not domain dependent which is a good feature for portability between the different fields of risk management application.

The proposed methodology uses the ontology of the PRIMA[1] project (supported by the European community) as a Generic Domain Ontology and populates it via an available corpus. A first validation of the approach is done through an experiment with Chemical Fact Sheets from Environmental Protection Agency[2].

Keywords— Information Extraction, Instance Recognition Rules, Instantiation, Ontology Population, POS tagging, Risk Management, Semantic analysis.

I. Introduction

Risk Management is as a rule assisted by decision support, which relies on a risk knowledge base (supposed to be or become a corporate memory) and a cognitive framework adapted to risk [1]. Our work focuses ~~is~~ mostly on the knowledge acquisition process related to the risk knowledge base. This acquisition process raises specific problems and difficulties: knowledge and expertise are sensitive, scattered, hidden or unclear; moreover, knowledge is highly specialized, although its implementation is very multi/inter-disciplinary. PRIMA represents the initial work of defining a Generic Domain Ontology, validated in industrial context, and kernel for further developments in the fields of ontology extension or content extension.

There is a variety of technologies involved into risk management systems that have been applied to support acquisition, creation, application and generation of organizational knowledge processes, such as: Databases and data warehouses, decision support system, expert systems, intelligent agents, data mining, ontologies, etc.

Our research focuses on ontology technology as the backbone to support the construction and the population of the Risk Knowledge Base, because of its power of expressivity and knowledge reuse capability.

Ontology is an explicit formal specification of a shared conceptualization of a domain of interest [3]. Ontology plays an important role in many kinds of applications especially in semantic web applications and knowledge management applications; it is widely used as a knowledge representation tool for domain knowledge. It defines concepts and relations between these concepts in order to represent knowledge in a specific domain. Ontology is well prepared by knowledge managers and domain experts. But it is a laborious, time consuming, tedious, cost-intensive and complex task to find concepts, to build relations and to add new instances in the ontology. Therefore there has been a growing interest in the (semi) automatic learning and populating ontologies.

In this paper we focus on ontology population. We propose a Semi-Automatic Ontology Instantiation approach that aims at enriching a Generic Domain Ontology by acquiring new instances of concepts from texts. Domain-specific ontologies are preferable since they limit the domain and make the applications feasible [4].

We have experimented our methodology in the domain of risk management by populating PRIMA ontology with instances through Chemical Fact Sheets from Environmental Protection Agency.

The rest of this paper is organized as follows: in Section II we present the State of the Art. In Section III we describe in details our approach of Ontology Instantiation. In Section IV An example experiment is detailed. Finally, in Section V we draw conclusions for further work and a future evaluation.

II. State of the Art

A. *Risk knowledge base within Risk Management and Business Intelligence:*

In every ontology approach, the definition of the sphere of work, the scope characterization, is compulsory to context understanding, requirement identification, usable sources recognition and functional analysis of users requirements. This is even more true in the area of risk management, which is a very generic problem, applying to all types of situations extending to:

- Varied levels of support, from individual commitment, business management to social problems. Here we aim to support the performance management of a company using

Jawad Makki : Université Toulouse 1, 2 rue du Doyen Gabriel Marty, 31042 Toulouse Cedex, email Jawad.Makki@univ-tlse1.fr
Anne-Marie Alquier : Université Toulouse 1, 2 rue du Doyen Gabriel Marty, 31042 Toulouse Cedex, email Anne-Marie.Alquier@univ-tlse1.fr
Violaine Prince: LIRMM-CNRS – Université Montpellier 2, 161 rue Ada, 34392 Montpellier Cedex 5, email prince@lirmm.fr

[1] PRIMA project : Project Risk Management, IST-1999-10193, 00-02.
[2] EPA : U.S. Environmental Protection Agency www.epa.gov/chemfact

risk, called management by risk. The general idea is that most business decisions are based on risk-taking - in the sense of opportunities as well as dangers.

- Every type of risk. For example, to consider whether the materials used in a new product have hazardous impacts of are environmentally friendly, many sources should be consulted, many of them being outside the company. Here we focus on a specific type of risk.

A risk knowledge base would capture as much knowledge as possible, capitalizing on all sources potentially useful, external or internal.

Internally, a risk knowledge base capitalizes the design, development, certification, operation and lessons learnt from the past.

But external knowledge is more useful is the field of strategic decision making [2] and Business Intelligence. Helping strategic decision makers is enabling them to operate more efficiently various data sources to get a better understanding of their organization and competitive environment.

It is then necessary to search for risk knowledge. Risk Management involves different organizations, at various levels. Knowledge is scattered in distinct systems and services. Extracting relevant knowledge is not just a raw data exchange with only the corresponding syntactic and semantic conversion issues well known in databases.

But this search should not be extensive: It is impossible and it would even be harmful to incorporate any type of risk for any type of organization in a risk knowledge base. It is necessary to focus on the needs related to specific situations.

Thus, knowledge acquisition cannot be fully automated; it should be rather guided by an expert, in a semi-automatic way. This expert is in charge of bringing together, from different sources, the domain-specific knowledge, in order to reuse it as a basis for risk and business intelligence, and to allow afterwards the simulation support in risk management. The knowledge-acquisition expert would in fact reengineer risk-oriented knowledge resources (databases as well as textual resources), subsequently mapping them to a central bone structure, an Enterprise Risk Management Mechanism.

Knowledge capture in heterogeneous, informal sources can be helped by the complex central ontology for classifying and managing risks provided by PRIMA. It includes domain, task, and problem-solving ontologies validated in several industrial contexts.

The semi-automatic acquisition process offers the following outputs:

- Meta-knowledge (all the classification methods are included in the knowledge base described by PRIMA).
- Risk identification (the ontology is a generic host structure, but evolution is possible with appends or changes)
- Detailed risk description (the cognitive framework ontology is a generic host structure, but evolution is possible with additions or changes).

B. Natural Language Processing (NLP) for Information and Knowledge Extraction

Natural Language Processing (NLP) has been largely addressed these last years as a proficient technology for text mining in order to extract knowledge: Relevant literature is so abundant that extensively referencing its items is a contribution by itself. Therefore, we will stick here to papers exploring text mining and NLP in applications related to Risk Management or to papers that inspired our model and methodology.

Two types of relationships between NLP and risk management can be found: Those dealing with risk definition in documents, assessing the difficulty of probability assignment to terms (natural language items) related to risk definition and management [5]. Probability is re-interpreted as a confidence value in a fuzzy logic approach to risk inference from a natural language description [6]. These two representative works in literature tackle a crucial issue in risk ontologies extraction from texts, documents or discourse (oral/written): Terminology is not as precisely defined as in domains like medicine or biology, risk assessment by experts in the shape of sentences does not naturally lead to an obvious formalization. Words and phrases are various, ambiguous, stylistic figures are numerous (metaphors, emphases, understatements). This drives researchers to reconsider knowledge extraction from texts as a more complex process than those described in the abundant biomedical terminology extraction literature (e.g. [7] which deals with one of the most typical aspects of knowledge extraction, Named Entities, and their insertion in a domain taxonomy). Researchers such as [8] have acknowledged the gap between textual input and knowledge as a structural pattern for a given domain: Authors suggest annotating corpora in order to provide clues for an efficient knowledge extraction. Annotation means a human intervention: It seems that more and more works recognize human judgment as an important element in the extraction process loop, a fact upon which our own approach is based (in section III). This is set up to reduce the liabilities of an automatic natural language processing extracting knowledge in such a difficult environment.

Our own approach benefits from existing NLP techniques in order to extract knowledge from natural language text. These techniques involve part-of-speech (POS) tagging in order to filter the most interesting categories, semantic networks to retrieve semantic relationships between phrases as concept instances, syntactic and semantic knowledge to build concept recognition heuristics applied to texts. More precisely we used TreeTagger [9] as a POS tagger providing a basic syntactic structure for text. For semantic relation extraction, we relied on WordNet [10], in order to expand some specific words with related terms. This expansion increases the chance of matching with other semantically similar terms and decreases the problem of linguistic variations. We also used it for the acquisition of synonyms. Last we relied on the predicative power of verbs to derive our concept recognition rules described in section III.

C. Ontology population with NLP techniques

Ontology population is the process of building the Knowledge Base. It consists of adding new instances of concepts and relations into an existing ontology. This process usually starts after the conceptual model of ontology is built.

As said in the introduction, building ontology and instantiating a knowledge base manually are a time-consuming and cost-intensive process. Therefore in recent years there have been some efforts to automate it. New approaches for (semi) automatic ontology population have emerged and considerably increased. These approaches are based on various techniques. ArtEquAKT [11]-[14] is a system that automatically extracts knowledge about artists from the Web, populates a knowledge base and uses it to generate personalized biographies. ArtEquAKT uses syntactic analysis to get the Part-Of-Speech and employs Semantic analysis to perform named entity recognition and extract binary relations between two instances. ArtEquAKT applies a set of heuristics and reasoning methods in order to remove redundant instances from the ontology.

LEILA [15] is an automatic approach that can extract instances of arbitrary given binary relations from natural language. LEILA uses a deep syntactic analysis and statistical techniques to learn the extraction patterns for the relation.

Reference [4] describes a pattern-based method to automatically enrich a core ontology with the definitions of a domain glossary. Reference [4] applies a method in the domain of cultural heritage. It is an automatic approach that extracts instances from semi-structured corpora (Art and Architecture Thesaurus) with the help of manually developed extraction patterns.

SOBA [16] is an information extraction system that automatically extracts information from heterogeneous sources (semi-structured data such as tables, unstructured text, images and image captions) and populates a knowledge base by using a standard rule-based information extraction system in order to extract named entities. These entities are converted into semantic structures with the help of special mapping declarative rules. SOBA addresses the problem of entity disambiguation by performing simple checks during instances creation.

These current approaches are based on various techniques; e.g. automated pattern recognition and extraction, statistic analysis, syntactic analysis, semantic analysis, mapping rules, etc. They differ from each other in some factors and have many features in common. Reference [17] defined the major distinguishing factors between ontology construction approaches. These factors are classified in the below categories "dimensions":

1) Elements learned: Concepts instances, relations instances.
2) Starting point: Domain ontology, Unstructured Corpus, Domain specific texts, Part-of-speech Tagger, Syntactic/Semantic analyzer, Manually engineered extraction patterns, additional resources (like WorldNet).
3) Learning approach: Statistical, logical, Linguistic based, Pattern extraction, Wrapper induction, combined.
4) Degree of automation: Manual, Semi-automatic (User Intervention), Cooperative, Full automatic.
5) The result: List of concept instances, List of relation instances, Populated Ontology.
6) Domain Portability: Limited, Domain specific, Fairly portable.

Risk management is multi domain, multi corpora with unstructured knowledge and sometimes with scarce knowledge. Machine learning approaches can not apply, so we needed to have a portable approach. Moreover as explained by [8] sole automatic approaches may misinterpret texts fragments which would be frequent as well as risky especially in the risk management domain. Last sticking with only one calculation method (symbolic, statistical) would deprive the system of the benefits of the other. Therefore, we have developed an appropriate method meant to be portable, semi-automatic and mixing several techniques. It is detailed in next section.

III. Our Approach

Our approach of ontology population is based on combined statistical, syntactic and semantic techniques. It starts with an initial generic ontology and a corpus of unstructured documents in a given domain, and produces a populated ontology as a result of the population process. The main steps of our approach are the following:

A. Annotation with part-of-speech tags
B. Semantic Relation Instances Extraction
C. Ontology Instantiation process

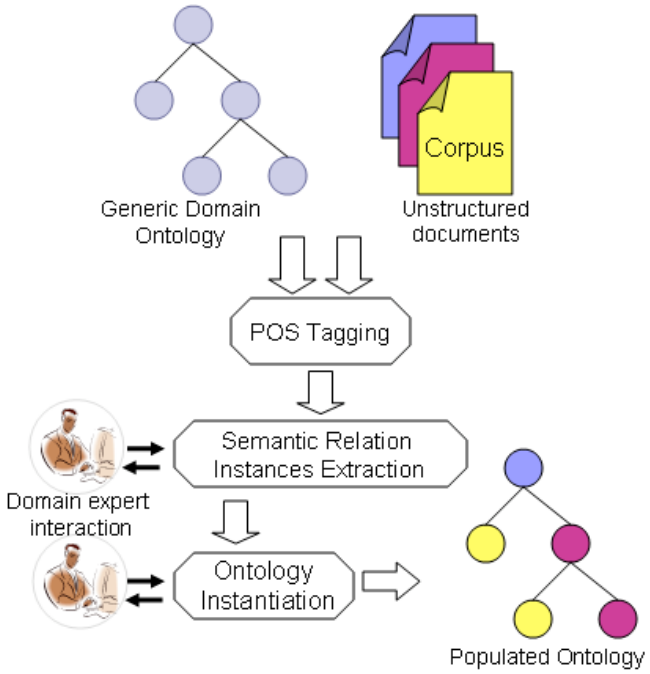

Fig. 1 Outline of the method

There is a loop between step B and C in which human interaction adjusts automatically extracted information and knowledge. The outline of the method is summarized in Fig. 1 and the three steps are detailed hereafter.

A. Annotation with part-of-speech tags

The corpus (i.e. any set of texts about risk considered as the source for knowledge extraction) is processed with TreeTagger. TreeTagger annotates texts with POS and lemma information. As a result, this step produced for each word w_i in the corpus, a string of POS tag or a syntactic category pos_i (e.g. NN for nouns, VB for verbs, JJ for adjective, etc…). These tags will be used as a filter to extract the frequent verbs

in next steps; it plays an important role also in the syntactic analysis.

Below is a sample output from TreeTagger for this sentence "*Prolonged dermal exposure to acetaldehyde can cause erythema and burns in humans*":

```
--------------- TreeTagger output -----------
Prolonged     JJ  prolonged
Dermal        JJ  dermal
exposure      NN  exposure
to            TO  to
acetaldehyde  NN  acetaldehyde
can           MD  can
cause         VV  cause
erythema      NN  erythema
and           CC  and
burns         NNS burn
in            IN  in
humans        NNS human
-----------------------------------------
```

B. *Semantic Relation Extraction / Semantic Relation Instances Extraction*

A semantic relation between two concepts may be expressed by a verb in natural language texts. Verbs represent an action or a relation between entities (concepts) in sentences. As a result, this step aims at generating semantic relation instances between concepts by extracting all frequent verbs from the POS annotated corpus in the previous step. These verbs are assumed to be associated with existing relations between two concepts from the ontology, which can be valuable for populating the generic domain ontology provided.

Let R_{ab} be a semantic relation between two concepts C_a and C_b of the ontology ($C_aR_{ab}C_b$). The idea is to construct from the annotated corpus, a list of verbs LV_{ab} associated to R_{ab} where each verb can link the two concepts C_a and C_b. this list will be validated by a domain expert.

The list of verbs LV_{ab} associated to R_{ab} is built by : 1) synonyms of R_{ab} generated by the lexical resource WordNet, 2) frequent verbs extracted from the annotated corpus (simple frequency counting) 3) Human interference by a knowledge manager or a domain expert where his role consists in validating the candidate set of verbs associated to R_{ab}.

In brief, this step of the method takes as an input the POS annotated corpus and the generic ontology and produces for each semantic relation R_{ab} between two concepts C_a et C_b of the ontology, a list of verbs LV_{ab} associated to R_{ab} where each v_i of LV_{ab}, v_i can semantically connect C_a and C_b.

$LV_{ab} = \{v_1, v_2, \ldots v_n\}$ where $\forall\, v_i \in LV_{ab}\ \exists\, C_a v_i C_b$

Example: For C_a = « Cause », C_b = « Risk » and R_{ab}= « provoke » the list of verbs LV_{ab} associated to R_{ab} is the following:

LV_{ab} = {provoke, evoke, cause, explode, result, fire …}

C. *Ontology Instantiation process*

From the list of verbs LV_{ab} semi automatically extracted from the corpus and for each verb V_{ab} of LV_{ab}, this step aims at identifying and extracting all triplets ($segment_i$,V_{ab},$segment_j$) from the set of sentences of the annotated corpus.

A triplet ($segment_i$,V_{ab},$segment_j$) is extracted from a sentence S that contains a verb V_{ab} of LV_{ab}. S is composed from a set of words w_i like S= $w_1 \ldots w_i\ V_{ab}\ w_j \ldots w_n$. $segment_i$ in triplet represents $w_1 \ldots w_i$ (i.e. the words left of the verb) and $segment_j$ represents $w_j \ldots w_n$. (i.e. the words right of the verb).

At a second phase, each extracted triplet is proposed to a syntactic structure recognition procedure; this procedure is based on a set of predefined Instances Recognition Rules. To initiate the ontology population process, these rules have been created manually by testing (we contemplate to automate this process in a further step with learning algorithms such as association rules if they prove to be numerous or if those we built up don't cover the problem). Rules can recognize a certain amount of linear words configurations. They are able to identify and generate an instance triplet (Instance_of_C_a, Instance_of_R_{ab}, Instance_of_C_b) from the extracted triplet ($segment_i$,V_{ab},$segment_j$).

However, these Instances Recognition Rules can be expanded with time through the addition of new rules in order to enhance the performance and the accuracy of the knowledge extraction method.

As a result, this procedure generates Instance triplets that have the form of ($Instance_a$,V_{ab}, $Instance_b$) where $Instance_a$ is an instance of concept C_a, $Instance_b$ is an instance of concept C_b and V_{ab} in an instance of relation R_{ab} that connect $Instance_a$ and $Instance_b$.

We distinguish in Table I some of the Instances Recognition Rules:

TABLE I
INSTANCES RECOGNITION RULES

Rule	linear words configurations	associated instances triplets ($Instance_a$,V_{ab}, $Instance_b$)
R1	$w_1 \ldots w_i\ \mathbf{V_{ab}}\ w_j \ldots w_k$	($w_1 \ldots w_i$,$\mathbf{V_{ab}}$,$w_j \ldots w_k$)
R2	$w_1 \ldots w_i$ DT $w_j \ldots w_k\ \mathbf{V_{ab}}\ w_1 \ldots w_m$ where DT = that	($w_j \ldots w_k$,$\mathbf{V_{ab}}$,$w_1 \ldots w_m$)
R3	$w_1 \ldots w_i$,$w_j \ldots w_k$,MD $\mathbf{V_{ab}}\ w_1 \ldots w_m$ where MD = can	($w_1 \ldots w_i$,$\mathbf{V_{ab}}$,$w_1 \ldots w_m$)
R4	$w_1 \ldots w_i$ NN1 $w_j \ldots w_k\ \mathbf{V_{ab}}$ WRB VV $w_1 \ldots w_m$ where WRB = when and NN1 (first noun)	(NN1 VV $w_1 \ldots w_m$,$\mathbf{V_{ab}}$, NN1 getNoun($\mathbf{V_{ab}}$))
R5	$w_1 \ldots w_i\ \mathbf{V_{ab}}\ w_j \ldots w_k$ CC $w_1 \ldots w_m$ where CC = and	2 triplets : ($w_1 \ldots w_i$,$\mathbf{V_{ab}}$,$w_j \ldots w_k$) ($w_1 \ldots w_i$,$\mathbf{V_{ab}}$,$w_1 \ldots w_m$)
R6	$w_1 \ldots w_i$ CC $w_j \ldots w_k\ \mathbf{V_{ab}}\ w_1 \ldots w_m$ where CC = or	2 triplets : ($w_1 \ldots w_i$,$\mathbf{V_{ab}}$,$w_1 \ldots w_m$) ($w_j \ldots w_k$,$\mathbf{V_{ab}}$,$w_1 \ldots w_m$)
R7	$w_1 \ldots w_i$ VBZ/VBP $\mathbf{V_{ab}}$ IN $w_j \ldots w_k$ where pos($\mathbf{V_{ab}}$) = VVN and IN = by VBZ = is ; VBP = are	($w_j \ldots w_k$,$\mathbf{V_{ab}}$,$w_1 \ldots w_i$)

The produced Instances triplets will be validated by a domain expert. Having the triplet in this form « $word_1 \ldots word_i$ + Verb + $word_j \ldots word_k$» facilitate the identification of instances of concepts/relations for the generic domain ontologies by the decision maker. Finally this validation instantiates the ontology and finishes the population process.

IV. EXPERIMENTS

In this section, we describe the application of our method for an example experiment in the domain of risk management.

In this experiment we use the ontology of PRIMA (the risk analysis reasoning model defined in PRIMA) as a generic ontology, and a corpus consists of 20 Chemical Fact Sheets (in English) provided by the Environmental Protection Agency.

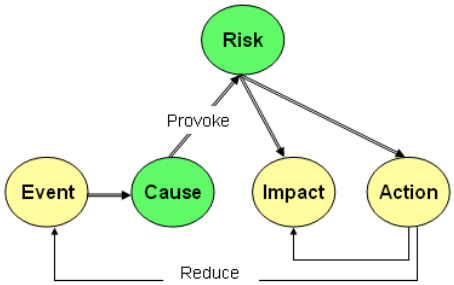

Fig. 2 Part of the Generic Ontology of PRIMA

The ontology of PRIMA contains a set of concepts describing risk and its insertion in a technical chain of work. Risk itself is described through 7 high level entities (or objects), the relations between those entities, plus relations with items external to risk (cost for example). Only two concepts and one relation were used for the experimentation in order to populate the causal chain of PRIMA and more specifically to instantiate the two concepts «Risk» and «Cause» and the relation «Provoke» that connects them.

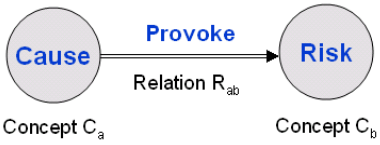

Fig. 3 the causal chain of PRIMA

After applying a POS tagging on the EPA corpus, we built the list of verbs LV_{ab} associated to Relation R_{ab} "Provoke" by getting the synonyms of relation "Provoke" and extracting all the frequent verbs associated to this relation from the EPA annotated corpus. Human intervention has validated the final list LV_{ab}.

In a second phase, we extracted all the triplets ($segment_i$,V_{ab},$segment_j$) and proposed them to the syntactic structure recognition procedure. This procedure generated 150 Instances triplets. 85% of these Instances triplets are evaluated as accepted instance triplets. Table II shows some results of our method.

Example 1 :

For V_{ab} = «cause» and for this entry "*dermal Prolonged exposure to acetaldehyde can cause burns and erythema in humans*", the Instances recognition rule R5 is applied and we get two instances triplets as following:

- *(Prolonged dermal exposure to acetaldehyde, Cause, erythema)*
- *(Prolonged dermal exposure to acetaldehyde, Cause, burns in humans)*

Example 2 :

For V_{ab} = «result» and for this entry "*Humans Toluene ingestion result in severe central nervous system depression*", the Instances recognition rule R1 is applied and we get an instance triplet as following:

- *(Humans Toluene ingestion, Result, severe central nervous system depression)*

TABLE II
SOME EXPERIMENT RESULTS

«Cause»	«Provoke»	«Risk»
Exposure to large amounts of chlorobenzene	cause	Adverse nervous system effects
Repeat exposure to nitrobenzene in air over a lifetime	Cause	cancer in animals
Prolonged dermal exposure to acetaldehyde	Cause	erythema
Prolonged dermal exposure to acetaldehyde	Cause	burns in humans
Methanol exposed to an open flame	explode	Explosion
Humans Toluene ingestion	result	severe central nervous system depression
Exposure to moderate amounts of chlorobenzene in air	Cause	Testicular damage in animals
Nitrobenzene	Cause	Adverse reproductive system effects
Chlorobenzene has potential	Produce	adverse reproductive effects in human males
Repeatedly breathing large amounts of toluene	Cause	permanent brain damage
……	……	……

V. CONCLUSION AND FUTURE WORK

In this paper, we presented an appropriate method for Semi-Automatic Ontology Instantiation from natural language text, in the domain of Risk Management. It's based on combined NLP techniques using human intervention for control and validation. First experimental results show that the approach reached 85 % of accepted Instances triplets. This percentage is satisfactory results encouraging us to go further:

- In populating other PRIMA concepts and relations within a given domain (here the chemical risk)
- In populating PRIMA generic ontology in other risk domain without extensive reworking.

Semantic relation extraction is not a domain dependent process and recognition rules are by definition domain independent (they are linguistic knowledge).

Since we rely so heavily on NLP, NLP limitations have a crucial impact on our method. For instance, POS disambiguation if not provided by the tagger could hamper recognition rules results («result» and «cause» are both noun and verb). Therefore, a deeper syntactic analysis than the one provided by TreeTagger, is investigated (we agree with LEILA authors and their choice of a real grammatical analysis).

However, our method ensures a real portability from a given domain to another if a generic ontology exists somewhere which is the case in risk management. It is flexible

(easily supports enhancement), useful for expert knowledge expression (it suggests word associations to risk experts which might give them a decision support).

REFERENCES

[1] Alquier A.M. & Tignol M.H., 2007. "Management de risques et intelligence économique", Economica. ISBN : 2717852522.

[2] Ansoff H.I., 1990. Implanting Strategic Management, Practice Hall.

[3] T. Gruber. Towards principles for the design of ontologies used for knowledge sharing. Int.J. of Human and Computer Studies, 43:907–928, 1994.

[4] R. Navigli and P. Velardi. Enriching a Formal Ontology with a Thesaurus: an Application in the Cultural Heritage Domain. In Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, pp. 1 – 9, Sydney, Australia, July 2006

[5] Hillson, D. 2005 "Describing probability: The limitations of natural language."Proceedings of EMEA, Edinburgh, UK.

[6] Huang, C.F.Risk 2007 "Analysis with Information Described in Natural Language ". In Computational Science, Proceedings of ICCS2007, Lecture Notes In Computer Science, Springer Verlag.

[7] Liang T, Shih PK 2005 Empirical Textual Mining to Protein Entities Recognition from PubMed Corpus, Proceedings of the10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain. Lecture Notes in Computer Science. Springer Verlag. Pp 56-66.

[8] Navarro B., Martínez-Barco P. and M.Palomar, 2005. "Semantic Annotation of a Natural Language Corpus for Knowledge Extraction" 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain. Lecture Notes in Computer Science. Springer Verlag.

[9] TreeTagger: a language independent part-of-speech tagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

[10] Miller, G. and McDonnell, J. S. 2003. "WordNet 2.0." A Lexical Database for English, Princeton University's Cognitive Science Laboratory. http://WordNet.princeton.edu

[11] S. Kim, H. Alani, W. Hall, P. Lewis, D. Millard, N. Shadbolt and M. Weal. Artequakt: Generating Tailored Biographies from Automatically Annotated Fragments from the Web. In Proceedings of Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM'02), the 15th European Conference on Artificial Intelligence, (ECAI'02), pp. 1-6, Lyon, France 2002.

[12] Alani H., Sanghee K., Millard E.D., Weal J.M., Lewis P.H., Hall W., and Shadbolt N., Automatic Extraction of Knowledge from Web Documents, In: Proceeding of (HLT03), 2003.

[13] Alani H., Sanghee K., Millard E.D., Weal J.M., Lewis P.H., Hall W., and Shadbolt N., Web based Knowledge Extraction and Consolidation for Automatic Ontology Instantiation, In: Proceedings of the Workshop on Knowledge Markup and Semantic Annotation at the Second International Conference on Knowledge Capture (K-CAP 2003), Florida, USA, 2003.

[14] H. Alani, S. Kim, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis and N.R. Shadbolt (2003), "Automatic Ontology-Based Knowledge Extraction from Web Documents", IEEE Intelligent Systems, 18(1), pp. 14-21.

[15] F.M. Suchanek, G. Ifrim and G, Weikum. LEILA: Learning to Extract Information by Linguistic Analysis. In Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, pp. 18 – 25, Sydney, Australia, July 2006.

[16] P. Buitelaar, P. Cimiano, S. Racioppa and M. Siegel (2006), "Ontology-based Information Extraction with SOBA", In Proceedings of the International Conference on Language Resources and Evaluation, pp. 2321-2324. ELRA, May 2006.

[17] M. Shamsfard, and A. Abdollahzadeh Barforoush. The state of the art in ontology learning: a framework for comparison. The Knowledge Engineering Review (2003), 18: 293-316 doi:10.1017/S0269888903000687